#4

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:
Bryant *et al.*
Appl. No. 09/514,955
Filed: February 28, 2000
For: **Clinical and Diagnostic Database**

Art Unit: 2771
Examiner: (to be assigned)
Atty. Docket: 1581.0590000/RWE

Claim For Priority Under 35 U.S.C. § 119(a)-(d) In Utility Application

Commissioner for Patents
Washington, D.C. 20231

Sir:

Priority under 35 U.S.C. § 119(a)-(d) is hereby claimed to the following priority document(s), filed in a foreign country within twelve (12) months prior to the filing of the above-referenced United States utility patent application:

Country	Priority Document Appl. No.	Filing Date
Great Britain	9904585.8	February 26, 1999

A certified copy of the listed priority document is submitted herewith. Prompt acknowledgment of this claim and submission is respectfully requested.

Respectfully submitted,

STERNE, KESSLER, GOLDSTEIN & FOX P.L.L.C.

Robert W. Esmond

Robert W. Esmond
Attorney for Applicants
Registration No. 32,893

Date: June 21, 2000

Sterne, Kessler, Goldstein & Fox P.L.L.C.
1100 New York Avenue, N.W.
Suite 600
Washington, D.C. 20005-3934
(202) 371-2600

P:\USERS\AMACWILL\RWE\1581\0590000\priorityclaim

The Patent Office

INVESTOR IN PEOPLE

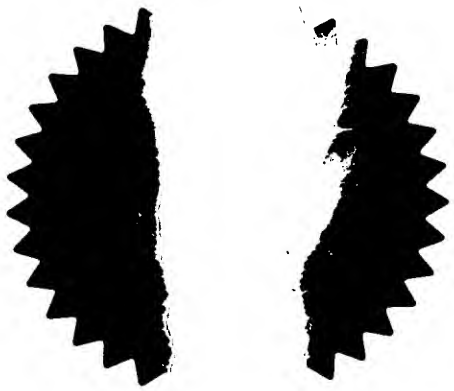
The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.



Signed Andrew Craig

Dated 16 MAR 2000

Patents Form 1/77

Patents Act 19
(Rule 16)

The Patent Office

1/77

01MAR99 E428925-1 D02000
P01/7700 0.00 - 9904585.8

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

1. Your reference — GWS\21265

2. Patent application number *(The Patent Office will fill in this part)* — 26 FEB 1999 — **9904585.8**

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*

   Gemini Research Ltd
   162 Science Park
   Milton Road
   Cambridge
   CB4 4GH
   U.K.

   Patents ADP number *(if you know it)*

   If the applicant is a corporate body, give the country/state of its incorporation

   7434640002

4. Title of the invention

   Clinical and diagnostic database

5. Name of your agent *(if you have one)*

   "Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

   MATHYS & SQUIRE
   100 Grays Inn Road
   London WC1X 8AL

   Patents ADP number *(if you know it)* — 1081001

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

Country	Priority application number *(if you know it)*	Date of filing *(day / month / year)*

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application	Date of filing *(day / month / year)*

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*

   a) any applicant named in part 3 is not an inventor, or
   b) there is an inventor who is not named as an applicant, or
   c) any named applicant is a corporate body.
   See note (d))

   YES

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form	
Description	34
Claim(s)	6
Abstract	-
Drawing(s)	-

10. If you are also filing any of the following, state how many against each item.

Priority documents	-
Translations of priority documents	-
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	-
Request for preliminary examination and search *(Patents Form 9/77)*	-
Request for substantive examination *(Patents Form 10/77)*	-
Any other documents *(please specify)*	-

11. I/We request the grant of a patent on the basis of this application.

Signature: Mathys & Squire — Date: 26 February 1999

12. Name and daytime telephone number of person to contact in the United Kingdom: GEORGE W SCHLICH — 0171 830 0000

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.

b) Write your answers in capital letters using black ink or you may type them.

c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.

d) If you have answered 'Yes' Patents Form 7/77 will need to be filed.

e) Once you have filled in the form you must remember to sign and date it.

f) For details of the fee and ways to pay please contact the Patent Office.

CLINICAL AND DIAGNOSTIC DATABASE

The present invention relates to a database containing information useful for clinical, diagnostic and other purposes, and relates in particular to a database containing genotype and phenotype information. The present invention also relates to methods of adding information to the database and to methods of identifying correlations within and between phenotypes and/or genotypes in the database as well as to other uses of the database.

It is recognised that most diseases can be correlated with geographical, environment, dietary, genetic and/or other specific contributory factors. Hence, much effort today is directed at identifying those contributing factors, and also those factors which may not directly contribute to these but are otherwise linked thereto and may be correlated with presence of disease for some other reason, so that even more accurate diagnosis of disease and pre-deposition to disease can be achieved.

It is known to select a group of individuals according to a particular criteria and carry out various tests including obtaining information such as genotype and phenotype to create a database of information concerning individuals conforming with the particular selection criteria chosen. Information in the database may then be used to identify causative factors or other factors related to incidence of or pre-deposition to disease. Following this strategy, it is known, for example, to carry out an analysis of the causes of the hypertension by selecting a group of individuals all of which are hypersensitive and then attempting to identify common genotypic or phenotypic characteristics amongst this group. When analysing the causes of the different disease, a different selected group of individuals is identified and may be subject of a separate analysis.

The present invention relates to a database and methods of maintaining the database and methods of use thereof which represents a new approach to obtaining correlations between phenotype and genotype as well as cross-correlation between phenotypes, and cross correlations between phenotypes and genotypes.

It is an object of the present invention to provide a database containing phenotype and also preferably phenotype information that can readily be used to obtain clinically and/or therapeutically and/or diagnostically useful information. A further aim is to provide a database of genotype and also preferably phenotype information which can readily be updated and expanded and adapted according to a wide ranges of uses proposed for the data within the database. A still further object of the present invention is to provide methods of obtaining clinically or therapeutically or diagnostically useful information from the data stored in the database of the present invention.

According to a first aspect of the invention there is provided a database comprising a plurality of records, each record containing phenotype information, and optionally sample information, for an individual, wherein:

> the phenotype information for the individual comprises at least osteoporosis related phenotypes, osteoarthritis related phenotypes, immune cell subtypes (such as Tcell subsets), metabolic syndrome/syndrome X related phenotypes, and hypertension related phenotypes; and
>
> the sample information for the individual comprises information relating to the location of a sample of tissue or of fluid from the individual.

The database is suitable for storage of records relating to a wide variety of different individuals, and is especially suitable for information relating to human individuals though it is equally suited for use with animal or other veterinary data, preferably mammalian data. The inclusion of sample information in the database enables users of the database to locate a sample of tissue or of fluid from the individual for further testing. This further testing might be to obtain additional phenotype information not previously tested from that tissue or fluid sample or it might be to confirm and possibly correct or update phenotype data already stored for a particular characteristic of that individual. The database is also suitable for correlation with other proprietary and public databases consisting of clinical

information, data on genomics, proteonomics, cell biology, immunology and biochemistry. Furthermore the database is interactive and allows cross correlation of key genotypes/haplotypes with key phenotypes to better understand the biology, and regulation of genetic, cell biological and humoral networks involved in complex diseases.

The type of tissue or fluid samples that can be stored in accordance with the invention are without limits. Typically, fluid samples that can readily be stored include urine, serum and saliva samples. Tissue samples that can readily be stored include skin, liver, heart tissue, bone, hair, muscle, kidney, tooth and faeces samples. Most of these tissue or fluid samples will contain DNA. Nevertheless, it is also an option for a separate sample to be stored containing DNA extracted from tissue of that individual. To enable easy location of the tissue of the fluid sample it is typical for the sample information to include the geographical location of the sample, for example the address of the storage institution, as well as the storage conditions and the storage reference number or storage identification number to enable identification and retrieval of the sample when needed.

Records in the database are preferred also to contain genotype information relating to the individual, such as one or more single nucleotide polymorphisms ("SNPs") in the DNA of the individual. Alternatively or additionally, the genotype information can comprise a record of actual or inferred DNA base sequence at one or more regions within the genome. Still further, the genotype information can comprise a record of variation between a specified sequence on a chromosome of that individual compared to a reference sequence; indicating whether and to what extent there is variation at identical positions within the sequence. The genotype information can yet further comprise a record of the length of a particular sequence or a particular sequence variant; such information being of use to investigate absence or presence of correlation between genetic variation and phenotype variation.

In this and related contexts, reference to genotype is intended to refer to genotype

or to haplotype or to both genotype and haplotype. In use of an example of the invention, SNPs from proprietary or public domain databases are added to and stored in the present database for the individuals. It is then possible to try to identify an association between one or more of these SNPs by correlation with one or more phenotypes stored in the present database. One method to achieve this is to search the DNA of an individual for one or more polymorphisms which are associated with a given risk trait, the polymorphisms being for example SNPs with allele frequencies of at least 20%, and which do not have linkage disequilibrium.

It is preferred that a large amount of phenotype information is recorded in the database for each individual, and also preferred that all or substantially all of this information is obtained via a single interview and/or examination or if necessary via numerous such sessions over a short time frame. The types of phenotypes stored can usefully include quantitative risk traits associated with chronic diseases, biochemical parameters, cell biological parameters such as cell surface markers and factors of cell growth, apoptosis and signal transduction, structural and humoral proteins and other biochemicals and metabolites.

In a preferred embodiment of the invention, the phenotype information recorded further includes thrombosis/fibrinolysis phenotypes, haemoglobinopathy related phenotypes and airways disease (asthma) phenotype. In this and related contexts, reference to phenotypes is intended to be a reference to data relating to at least one phenotype and typically more than one phenotype of the nature indicated. Additional phenotype information used in still further preferred embodiments of the invention relates to the phenotypes: atopy/eczema, lung function, IgE, psoriasis, acne, skin cancer and moliness of skin.

Other information that may be included in the category of phenotype information that can be included in the database comprises information relating to quantitative traits related to cognition, dementia, parkinson's disease and intelligence, history of adverse drug reactions and history of substance abuse/addictive behaviour.

It is thus apparent that the database of the invention may hold information on phenotypes in a hitherto unmatched number of categories. This extensive breadth of information in specific embodiments of the invention contributes to the uniquely valuable information that can be extracted therefrom in the various applications of the database described below.

Still further optional areas of phenotype information that are include in the database relate to: lifestyle - such as alcohol, tobacco, diet, exercise - , dietary history, medication history and family history of disease.

The sample information may additionally include contact information so as to enable the individual whose data is already in the database to be contacted and recalled for further testing.

It is an advantage of having the sample information that data in the database can be checked, corrected and/or expanded by further testing of the tissue or fluid samples that have been stored for each individual. In the case of an unusual value being recorded for a particular phenotypic characteristic, a tissue or fluid sample can be retested to confirm the information in the database. Whilst it is believed that the phenotype stored in the database will be sufficient to enable a wide range of uses of the data, it is envisaged that some particular investigations will call for phenotype information that has not yet been tested for individuals in the database, or has not been tested in the manner required for a particular investigation. In these circumstances it is particularly advantageous that the tissue or fluid sample can be recalled and tested to add in the required additional phenotype information to that phenotype information already present in the database. The further testing of stored material in this way is considerably more convenient and efficient than trying to locate individuals that have been included in the database and arrange for further testing of missing phenotype information in person.

In a database of the invention, phenotypic data are generally maintained for each individual within the database with most data being associated not only with an

individual, but also with a particular timepoint. Some physiological results vary over time and are valid in relation to each other only if collected at the same timepoint.

Stored material (DNA, Serum and Urine) is preferably maintained for each individual, for each visit. Additional phenotype data may be collected by performing assays on stored material, which will not deteriorate appreciably, even over several years. There is therefore the potential to expand the phenotype within the database of the invention, even if the assays are not carried out at the time of the visit. It is also possible to expand the phenotype by conducting questionnaires, interviews or other measurements, if the results are not expected to vary over time, or else vary predictably. This can include a) historical medical data, b) family history and c) drug usage.

There is also the option of collecting longitudinal data by having the individual return for a repeat visit. In this case, all the time-sensitive results are distinctly recorded within the database, which permits another dimension of analysis (time, or ageing) to be carried out. Some measurements from repeat visits would not necessarily be time-dependent and could be analyzed against results collected at earlier visits. Also, new technologies are brought in from time to time and can be used to "top-up" the phenotype. For straightforward analyses of a single outcome phenotype against the genetic background (which does not vary over time), it does not matter that these additional phenotypes are collected over a period of years, and this method is validly used to expand the database phenotype by a managed programme of revisits.

In a further embodiment of the invention, there is provided a method of integrating (a) information either in the private or public domain on genomic, proteonomics, cell and molecular biology and /or immunology with (b) information on the database of the invention, which information is collected on the patient population, and determining if there are any correlations between them.

In a second aspect of the invention, there is provided a method of adding information to the database of the invention, comprising:

1. identifying an individual not yet included in the database;

determining phenotype information for the individual that comprises at least osteoporosis related phenotypes, osteoarthritis related phenotypes, immune cell subtypes (such as Tcell subsets), metabolic syndrome/syndrome X related phenotypes, and hypertension related phenotypes;

optionally determining genotype information for that individual;

optionally determining sample information for the individual that includes information relating to the location of a sample of tissue or of fluid from the individual; and

creating a record in the database to hold the phenotype and optionally genotype and/or sample information for the individual;

or

2. identifying an individual already included in a record in the database;

using sample information in the database to obtain a tissue or fluid sample for the individual;

testing the sample, thereby determining genotype or phenotype information for the individual; and

adding or confirming or amending or updating information in the record for the individual.

The method of the second aspect of the invention represents improvement over the operation of prior art databases, in that the information stored in the database of the present invention can continually and without limit be expanded and updated and, if need be, corrected. The information in the database of the present invention does not reach a point at which it needs to be discarded and a new database started. Instead, the information can be obtained and amassed in a cumulative way so that the database is forever becoming more useful and more accurate for obtaining clinically or therapeutically or diagnostically useful information. It is particularly preferred that the information stored in the database of the invention is obtained from individuals who have not been selected according to any particular genotype or phenotype characteristic. That is to say, whereas in the prior art a cohort of individuals might have been selected for use in a genotype and phenotype database because they all had low bone mineral densities, the individuals included in the database of the present invention are not selected in this way. Instead, genotype and phenotype information from all and any individuals may be included in the database.

A disadvantage of prior art databases was that the cohort of individuals selected, for example, for an investigation into bone mineral density and the factors affecting bone mineral density would not be suitable for a separate investigation into, say, the effect of diet on blood pressure. The database of the present invention does not suffer from this disadvantage because the individuals in the database of the present invention have not been selected with any one particular clinical investigation in mind and are advantageously suitable for use in substantially all such investigations.

Further aspects of the invention relate to uses of the information contained in the database of the invention. Accordingly, a third aspect of the invention provides a method of identifying a correlation between phenotype information and genotype information comprising:

selecting a phenotype characteristic;

identifying a plurality of records from the database of the invention for individuals that comply with the selected phenotype characteristic;

determining if presence of the selected phenotype characteristic is correlated with presence of any genotype characteristic in the genotype information for records in the database.

A fourth aspect of the invention provides a method of identifying a correlation between phenotype information and phenotype information comprising:

selecting a phenotype characteristic;

identifying a plurality of records in the database for individuals who comply with the phenotype characteristic;

determining if presence of the selected phenotype characteristic is correlated with another characteristic of phenotype information for records in the database.

More specifically, the method can comprise identifying correlation between presence of the selected phenotype characteristic and two or more separate characteristics of phenotype information for records in the database

A fifth aspect of the invention provides a method of identifying a correlation between genotype information and genotype information comprising:

selecting a genotype characteristic;

identifying a plurality of records in the database for individuals who comply with the genotype characteristic;

determining if presence of the selected genotype characteristic is correlated

with another characteristic of genotype information or records in the database.

In use of the invention, there is provided a method of allocating priority to a candidate gene or locus, proposed as a drug target for treatment of a disease, the method comprising:-

calculating, from data on a database according to the invention, the specificity of the candidate gene or locus for the disease;

comparing (i) the association of the disease with clinical risk traits related to the disease, to (ii) the association of the disease with other clinical risk traits unrelated to the disease, but representing significant side effects; and

hence calculating a likely therapeutic index of drug candidates acting on that gene or locus.

For a top priority gene, the information on the database is used for correlating genotype with clinical risk traits, and with associated biochemical and cell biology phenotypes. This can give valuable information on the targets and mechanisms of action, and the biochemical pathways.

In a further general use of the invention, there is provided a method of analysing the relation between a genotype and a phenotype, comprising

selecting a phenotype characteristic;

identifying a plurality of records complying with that characteristic;

using environmental and age-related data in the database to eliminate the effects of age and environment on variations in phenotype; and

hence calculating from the database whether and if so to what extent the phenotype is correlated with a particular genotype.

In a further example of the invention in use, there is provided a method of determining the capacity and specificity of a genetic marker to detect and quantify normal variations in healthy and affected populations for a selected risk trait, comprising:-

assaying a sample in the database for the marker levels, in both healthy and affected subjects; and

quantifying the association of the clinical trait with the marker level and other selected phenotypes, in unaffected and affected subjects.

Another use of the invention lies in a method of predicting the response of patients to a selected drug therapy in a clinical trial, comprising:-

selecting a proposed clinical population for the trial;

using data on the database to stratify the clinical population by high associations of metabolism/absorption both with genotype and/or with associated biochemical and cell biology phenotypes; and

hence allowing definition of the best dose regimes and dose forms/drug delivery systems;

so as to predict and/or allow for absorption and/or metabolism of the drug by patients in the clinical population.

A yet further example of the invention in use provides a method of predicting response to a proposed drug therapy, comprising:-

using the database to select a clinical population by constructing haplotypic profiles, with strong associations with defined clinical traits and biochemical phenotypes;

using the database, and the twin resource, to eliminate the effects of age and environment in the clinical population;

hence providing criteria to predict response to the drug and variation in response to the drug, and optionally to define a sub-group of the clinical population or of the general population most susceptible to the drug being studied.

Twins are useful for controlling quantification of the impact of environmental factors on disease risk and are suitable for inclusion in a database of the invention. Identical twins share the same genes so any difference in a clinical measurement within an identical twin pair must be due to environmental factors or measurement error. By studying sufficient numbers of identical twins and measuring relevant environmental factors one can quantitate the impact of the environmental on clinical measurements.

Also, twins can be identified who are discordant for an environmental exposure. For example by examining fat mass where one twin from sufficient numbers of subjects where one identical twin of a pair smokes and the other does not one can quantitate the impact of smoking on obesity (Samaras et al Int J Obesity 1998). This can be made more sophisticated by doing such an analysis in twins who are concordant or discordant for other environmental factors, for instance exercise level. If the quantitative impact of various environmental factors is also known then one should be able to integrate that information into a multivariate model, along with candidate gene or candidate loci data, to identify gene-environment interactions.

Twins are followed prospectively and have further phenotypic data collected and

also further DNA, serum, urine or tissue samples collected.

Samples taken from twins at any one clinical visit are stored to be used at any future. These can be reanalysed for new biochemical or serological analytes and related to historical clinical and genetic data. Moreover, DNA is stored and can be retrieved for further genetic analysis as required. Lymphocytes cells are frozen and stored for future immortalisation to allow an 'infinite' DNA resource.

Phenotypes relating to many clinical diseases (either their presence or absence or the risk of these diseases) in the twins novel correlations between phenotypes can be identified that could not be so if the data collection was solely focused on a more limited phenotype set. This is carried out by various forms of correlational and cluster analysis to identify novel relationships between quantitative traits relating to broad disease areas. For instance relating phenotypes in anxiety and depression to those involved in diseases such as diabetes, osteoporosis, immunity, coagulation, may identify novel new disease entities that will be useful for

clinical diagnosis;

design of clinical trials;

targeted therapeutic intervention;

identification of new disease targets for drug discovery;

identification and validation of new molecular targets for drug discovery programmes; and

identification of patient populations most susceptible to chronic illnesses and hence to therapy.

There now follows description of specific embodiments of the invention for the purpose of non-limiting exemplification thereof.

EXAMPLE 1

MAKING A NEW ENTRY IN OR
AN ADDITION TO THE DATABASE

1. Initial Telephone Interview

The below-described protocol is followed to make a new entry in the database or to make an addition (or other change) to existing data.

The first stage is a telephone interview with one twin to request the following information:

Date of Birth
Address
Sex
Menopausal Status
Zygosity
Any serious illness or clinical conditions
How the interviewee heard about the study
Why the interviewee wishes to participate

The responses are recorded in an administration database and are used when calling subjects for interview as and when required.

2. Arrangements for the Study Day

Any individual who has had the initial interview may be called. Alternatively, as and when requirements for particular kinds of twin arises (e.g. sex, age) the database is interrogated and details of twins with the relevant profile are flagged out of the system – the example is thus written for the case that twin data is being added, though the same protocol is used for non twin data.

3. The Study Day

The following routine tests are carried out on each twin:

Fasting blood tests
Urine Tests
Anthropometric Measurements
Blood Pressure
Arterial Distensibility
DEXA Scanning: bone density and body composition
Muscle Strength : leg extensor power rig
Heel Ultrasound Scan
Spirometry
Electrocardiography
MRI Scans
X-Rays

Occasionally, other tests will be added for a particular study. A checklist is compiled for each test, completed as the interview progresses.

Questionnaires are also administered to the twins. Some during the study day, some which are sent out with the appointment letter and others provided as "homework" to complete after the visit day for sending in to the unit at a later date. The questionnaires contain a large number of questions on family history, medical history, current status and physical findings. Prospective questionnaires are required on certain clinical topics. In such cases, twins are given questionnaires to complete at home after the visit.

4. Processing Blood and Urine Samples

The following samples are taken:

Time 0 Glucose Tolerance Test (GTT)

30 ml clotted sample (3 x 10 ml tubes brown top)
40 ml EDTA (4 x 10 ml purple top)

2 ml fluoride/oxalate tube (grey top)

Time 120 after GTT (if done)

10 ml clotted sample (1 x 10 ml plain tubes brown top)
2ml fluoride/oxalate tube (grey top)

4.1 Clotted Samples

Clotted samples for serum are spun at 3000 rpm in a suitable centrifuge for 10 minutes after standing for 2-4 hours.

Time 0 samples

1 x 500 microlitre sample for routine biochemistry
12 x 1.5 ml cryotubes with green tops (approximately 750 microlitres/tube)
1 x 300 microlitre sample for sex hormones (as requested)

b. Time 120 samples

4 x 1.5 ml cryotubes with green tops (approximately 750 microlitres / tube)

4.2 EDTA samples

These samples are for DNA extraction.

a) the sample is spun at 3,000 rpm for 10 minutes in a clinical centrifuge;

b) the buffy coat (the leucocytes, a yellowish layer of cells on top of red blood cells) is removed and pooled into a 15ml conical tube;

c) 0.9% saline is added to fill the tube and resuspend the leucocytes. If there is a time delay, the sample can be stored at 4°C for up to 48 hours;

d) the sample is spun at 2,500 rpm for 10 minutes at 4°C;

e) the buffy coat is again removed as cleanly as possible leaving behind any red cells, the sample is suspended in cold red cell lysis buffer and left for 20 minutes at 4°C;

f) the sample is spun again at 2,500rpm for 10 minutes. If a pellet of unlysed red cells remains lying above the leucocytes, the treatment with red cell lysis buffer is repeated;

g) the leucocyte pellet is resuspended in 1 - 2ml 0.9% saline;

h) the DNA is liberated by the addition of 3ml leucocyte lysis buffer - the tube is capped and gently inverted several times, when the liquid will become viscous with DNA. The sample should be handled with care to avoid shearing and damage to the DNA;

i) proceed to DNA extraction.

4.3 FLUORIDE/OXALATE SAMPLES

The Time 0 and 120 tubes are sent directly to the Chemical Pathology laboratory.

4.4 URINE SAMPLES

Two aliquots are stored in 1.5ml cryotubes (750ul/tube yellow tops).

4.5 LOGGING LABELLING AND STORAGE

4.5.1 LOGGING AND LABELLING

All samples are given a unique laboratory code number and logged into the Twin Unit laboratory database. This number is used on all labels to identify all samples for a twin subject for a given visit date.

4.5.2 STORAGE

Those samples for immediate testing have no special storage;

Serum and urine samples which are stored at -45°C for batched assays will be given a unique freezer location code.

4.6 SENDING SAMPLES FOR ASSAY

Appendix 1 shows the scheme for the handling/testing of blood samples.

4.6.1 DAILY

The 1 x 500ul routine biochemistry sample (see 5.2.1. a)) is placed in the Chemical Pathology request bag, with the 0 and 120 minute fluoride/oxalate samples. A "Twin Label" (see SOP 2) is attached to the bag, which is taken to Chemical Pathology for routine biochemistry. If sex hormone estimations are to be carried out the extra tube is included. The assays are completed on the day of the sampling, or after storage overnight. If the samples are tested next day, the fluoride/oxalate samples are spun and the clot discarded before storage.

4.6.2 OTHER

All other research assays are sent to other laboratories and carried out as required from the frozen serum and urine samples (see 5.3.2. b)).

4.7 ASSAYS

The following assays are carried out.

4.7.1 ROUTINE BIOCHEMISTRY

sodium
potassium
chloride
bicarbonate
urea
creatinine
total protein
albumin
phosphate
total calcium
total bilirubin
alanine amino transferase
total alkaline phosphatase
magnesium
uric acid

4.7.1 GLUCOSE

From fluoride/oxalate samples

4.7.2 LIPIDS

Measured in one aliquot after storage at -45°C:

triglycerides
high density lipoproteins

apolipoproteins A1

apolipoproteins B

lipoprotein A

cholesterol

4.7.3 INSULIN

Measured in one aliquot after storage at -45°C.

4.7.4 SEX HORMONES

Measured in one aliquot:

follicle stimulating hormone (measured on the day of visit)

testosterone

Measured in one aliquot after storage at -45°C (if required):

sex hormone-binding globulin

dehyroepiandrosterone

4.7.5 BONE SPECIFIC MARKERS

Measured in one aliquot after storage at -45°C:

vitamin D binding protein

Measured in one aliquot after storage at -45°C:

bone-specific alkaline phosphatase

4.7.6 VITAMIN D METABOLITES/BONE FORMATION MARKERS

Measured in one aliquot after storage at -45°C:

1,25 (OH) vitamin D

Measured in one aliquot after storage at -45°C:

Parathyroid Hormone (PTH)

Measured in 2-3 aliquots after storage at -45°C:

25 (OH) vitamin D

4.7.8 THYROID FUNCTION

TSH
FT3
FT4

4.7.9 LEPTIN

4.7.10 URINE

Measured in one aliquot after storage at -45°C:

calcium
creatinine
deoxypyridinoline (Type 1 collagen crosslink)

4.7.11 EXTRA TESTS

Extra test may be done for special protocols.

5. Use of sample taken from individual already tested

The above description applies to the case that an individual is newly added to the database of the invention. The tests described, whether just one or any combination thereof, carried out on samples obtained from the individual are also repeatable using those samples to correct or confirm existing data or to carry out a test for the first time.

EXAMPLE 2

MAKING A NEW ENTRY IN OR AN ADDITION TO THE DATABASE

As an alternative or addition to the protocol of Example 1, the following phenotypic data are obtained for the record of an individual on the database.

Primary

The individual is tested for information relating to the following, referred to as "primary", phenotypes:-

Osteoporosis related phenotypes

Bone ultrasound
Bone density (total and regional)
Bone remodelling markers
Calcitropic hormones
Vitamin D and metabolites
Bone size
Postural stability
Fracture History

Osteoarthritis related phenotypes

Scores based upon x-ray (radiological, hands, knees, and hips on all twins > 40yrs)
Muscle strength
Disc Degeneration Indices (by Magnetic Resonance Imaging)
Serological markers of Inflammation

Immune cell subtypes (Tcell subsets)

Immunoglobulins

Dynamic responses of immune cells to stimuli

Metabolic Syndrome/Syndrome X related phenotypes

Fasting insulin and glucose

Insulin and glucose 120 minutes post glucose load

Leptin

Lpa

HLDL, Chol, Trigs, ApoB, ApoA

Obesity (total and regional, by direct measures of adiposity)

Hypertension related phenotypes

Cardiac Disease (heart chamber and size and dynamics on echocardiography)

Arterial tonometry and distensibility,

Central arterial pressure, pulse wave velocity

Thrombosis/fibrinolysis phenotypes

Haemoglobinopathy related phenotypes

Airways Disease (Asthma)

Atopy/Eczema

Lung Function

IgE (specific)

Psoriasis

Acne

Skin Cancer

Moliness of Skin

Quantitative traits related to Cognition, Dementia, Parkinson's disease and intelligence

History of adverse drug reactions

History of substance abuse/addictive behaviour

Secondary:

The individual is optionally tested for information relating to the following, referred to as "secondary", phenotypes:-

Lifestyle

- Alcohol
- Tobacco
- Diet
- Exercise

Comprehensive dietary history (validated)

Medication history

Family history of disease

EXAMPLE 3

The database of the invention can be used in the following applications.

A. Prioritisation f candidat genes, and Validation of high valu drug targ ts

These applications are relevant in cases where:-

several genes and/or gene regions are known which may contribute towards clinically significant risk traits; and

it is desired to prioritise one or a small number of these drug targets, and validate them.

This is achieved in the following ways.

The database including its twin resource is used to eliminate the effects of age and environment on variations in phenotypes.

The database is used to locate the gene(s) with a role in a given risk trait(s), sequence the gene(s) and identify mutations in the gene(s).

Polymorphisms with allele frequencies of at least 20% and with no complete linkage disequilibrium are selected to eliminate redundancy.

Each remaining polymorphism can be tested for association with selected phenotypes using a mean effect model.

Those phenotypes with high association with a given gene or locus can be identified - these phenotypes could be: other clinical risk traits, cell biology markers or surface receptors, circulating plasma proteins and immunoglobulins, clinical chemistry markers, circulating levels of hormones and other metabolites.

Each polymorphism can be analyzed for linkage to the candidate gene using single and multi-point linkage analyses.

The contribution of several candidate genes towards clinical risk traits, which contribute significantly to the disease can be quantified.

For the top priority gene(s), the information on the database is used for correlating genotype with clinical risk traits, and with associated biochemical and cell biology phenotypes. This gives valuable information on the targets and mechanisms of action, and the biochemical pathways.

The database is used to calculate the specificity of the candidate gene or locus, and hence the likely therapeutic index of drug candidates acting on that gene or locus, by comparing the association with clinical risk traits related to the disease, to other clinical risk traits, unrelated to the disease, but representing significant side effects.

B. Screening and validation of new genotype or phenotype markers

These application are relevant to the case that a several new markers have been identified (such as genetic, protein or other biochemical and/or cell biological markers) and it is desired to investigate both their clinical significance and specificity. Assay methods may already be known for the markers, though it may be desired to quantify the heritability of the markers, and to prioritise and validate them, so as to decide which ones to develop.

The database of the invention can be used to determine the heritability, and prioritise and validate the markers by:

- using the database, and the twin resource, to eliminate the effects of age and environment on variations in marker levels.

- assaying the blood/urine samples in the database for the phenotypic marker levels, in both healthy and affected subjects.

- locating the gene(s) with role in given risk trait(s), and sequencing the gene(s) and identifying mutations in the gene(s).

- selecting polymorphisms with allele frequencies of at least 20%, and with no complete linkage disequilibrium to eliminate redundancy.

- testing each remaining polymorphism for association with selected clinical traits and marker levels using a mean effect model.

- quantifying the association of the gene (locus) with the clinical trait and marker level.

- quantifying and comparing associations with other clinical traits

- hence quantifying the specificity of the marker to detect the clinical trait.

In the case that there are no candidate genes, the database can be used to prioritise and validate the markers by:

- assaying the blood/urine samples in the database for the marker levels, in both healthy and affected subjects.

- quantifying the association of the clinical trait with the marker level and other selected phenotypes, in unaffected and affected subjects.

Thus for a given marker, the database can be used to determine its capacity and specificity to detect and quantify normal variations in healthy and affected populations for selected risk traits. A decision can then be taken as to whether and how to develop the marker(s).

C. Accelerated and more effective clinical development

Selection of clinical indications for investigation

These applications are relevant where there is a lead candidate in development, or

a product on the market, which is desired to be put into clinical testing. It may be desired either to define the best clinical indication(s) or, for a selected indication, to identify patient populations which would best respond to the drug therapy. In these circumstances, the database of the invention can be used to assist in this analysis by:

- using the database, and the twin resource to eliminate the effects of age and environment on variations in drug response.

- constructing haplotypic profiles, with strong associations with clinical traits and biochemical phenotypes.

- hence prioritising the clinical traits and the indications in which the drug is likely to be effective

- defining methods for stratifying clinical trial populations for any clinical trait by haplotype and/or by phenotype.

- defining selection and exclusion criteria for patient recruitment, leading to better design of clinical trials, speedier clinical trials and an ability to achieve significant results on smaller patient populations.

- defining biochemical and cell biological profiles for patient selection and hence obviating the need for haplotyping, and the associated logistics, legal and ethical problems.

Selection of the most appropriate dose regimes and drug delivery systems

The absorption metabolism (pharmacokinetics) and even mechanism of action (pharmacodynamics) of drugs is affected by several enzymes, and this leads to large variations in the response by patients to drug therapies. The database of the invention can help to optimise dosage regimes and dose forms by:

- using the database, and the twin resource, to eliminate the effects of age and environment on variations in absorption, metabolism and mechanism of action.

- sequencing the gene(s) and identifying mutations in the gene(s).

- selecting polymorphisms with allele frequencies of at least 20%, and with no complete linkage disequilibrium to eliminate redundancy.

- testing each remaining polymorphism for association with selected absorption, metabolic phenotypes and with associated biochemical and cell biology phenotypes using a mean effect model.

- stratifying the clinical populations by high associations of metabolic/absorption and other phenotypes both with genotype and/or with associated biochemical and cell biology phenotypes.

- hence allowing definition of the best dose regimes and dose forms/drug delivery systems.

Clinical trials

The database of the invention can be used to provide, in connection with clinical trials:

- prediction on how patient populations will respond to drug therapies.

- better designed phase 1 Studies - the stratification of a volunteer population by pharmacokinetics and pharmacodynamics could give far better data, and indeed more than one dose regime and dose form could be tested so as to provide the best profile of the drug for a defined patient group. It might even be worth testing more than one candidate drug.

- better designed phase 2 Studies - such data can be used for phase 2 studies against comparators. Because the candidate drug, dose regimes and dose forms have been optimised during phase 1, phase 2 studies could be performed with far better exclusion criteria, would stand a far better chance of showing important differences, (important for studies with large placebo effects), and would need fewer patients recruited. This would reduce the time needed for the studies.

- better designed phase 3 and phase 4 Studies - the genotyping and phenotyping results from phase 2 studies can be further refined for phase 3 studies - which are in much larger patient populations, and consume the most time and money. The benefits are the same as above, but far larger. The same applies for the design of phase 4 (post marketing), when data on even larger patient populations are available.

- patients would have more appropriate and possibly individualised dosage and treatment regimes.

- specific dose forms and drug delivery systems could be developed for defined patient populations.

- information on responders and non-responders would minimise toxicity.

- pharmacoeconomics - better data to support demands for regulatory approvals and pricing and reimbursement. (better defined patient populations, better efficacy of treatment/lower treatment costs for health authorities).

- differentiating claims over competitive products.

- post marketing clinical studies - as more data is available on a wider patient population, and there are more side effects, then more refined genotyping/ phenotyping could define parameters so as to enable the drug to stay on the market. The database could be used to correlate data on disease parameters with

data on risk traits.

D. Epidemiological studies

These application apply where it is desired to carry out epidemiological studies on the effects on drug therapy, vaccination or an environmental pollutant. The database of the invention can help to define the population for the design by:

- using the database, and the twin resource, to eliminate the effects of age and environment.

- defining clinical populations by constructing haplotypic profiles, with strong associations with defined clinical traits and biochemical phenotypes.

- hence providing criteria to explain the variation in response, and define the groups most susceptible to the factor being studied.

E. Studying complex diseases

During clinical studies on unselected populations, several clinically significant risk traits may be identified, and associated with the complex disease.

By using the database of the invention and associated databases covering: genomics, proteonomics, cell biology and biochemistry, it is possible to:

- analyze the interaction of genes with other genes, and with proteins and other metabolites

- determine genetic and non-genetic networks (e.g metabolic).

- hence determine the metabolic pathways and regulatory mechanisms.

- validate high value molecular targets.

EXAMPLE 4

Samples used in connection with the database and their respective sample information are processed as follows.

Frozen samples (DNA, serum and urine, or any other clinical material) are transported from the collection centres to the database manager, using an approved courier. Samples arrive along with an electronic file and a printout of what has been sent. This should include a consignment number assigned by the collection centre, Study number (and checksum), DOB, lab reference, zygosity (in the case of twins), family number (if applicable) and volume and concentration if this is available.

Samples are logged into the database by manual or electronic entry of accompanying information. An aspect of the database is a sample tracking system, which allocates, and tracks the physical whereabouts of the samples within the database freezers. For security, each sample is stored in freezers in at least two separate buildings. Aliquots of samples may be measured, divided, diluted or concentrated by conventional means as is required for subsequent analysis. Where necessary the location of processed aliquots is allocated and tracked by the sample tracking aspect of the database.

DNA samples are subjected to any of a number of established laboratory procedures for the determination of actual or inferred DNA base sequence at regions within the human genome. The regions may be of any size (> one nucleotide) and anywhere within the genome. They are each usually defined by prior knowledge of the base sequence of a part or the whole of the region in at least one human individual.

Where the purpose of determining DNA base sequence is to discover

novel/unpublished sequence in one or more human individuals, the determined sequence is entered into an aspect of the database. The method of entry and format of sequence depends on the method used for determination. The sequence is stored for reference and such further data analyses as may be required. An example of further analysis could be to identify gene coding sequence.

Where the purpose of determining DNA base sequence is to discover sequence variation between two or more chromosomes (in one or more individuals) at identical positions within the sequence, the information pertaining to the sequence variation is entered into an aspect of the database. The method of entry and format of information depends upon the method used for the determination. The sequence variation is stored for reference and such further data analyses as may be required. An example of further analysis could be to investigate the effect of the sequence variation on gene coding sequence.

Where the purpose of determining or inferring DNA base sequence is to identify and record the particular sequence variations (genotypes) in one or more individuals, the genotypes are entered into an aspect of the database. The method of entry and format of genotypes depends on the method used for the determination. The genotypes are stored for reference and such further data analyses as may be required. An example of further analysis could be the identification of an association between hypertension and an identified locus.

Whether the genetic information be a length of sequence, a particular sequence variant, or genotypes in one or more individuals, in conjunction with the phenotype information it is able to be used (in a myriad of ways) to investigate the absence or presence of correlation between human genetic variation and human phenotype variation. Any combination of genotypes and phenotypes that resides within the database can be available for analysis. Such correlations are either directly or indirectly indicative of a causal relationship between the genetic region/s and the phenotype/s, under investigation. The utility of the database is to confirm, refute, or discover such correlations.

The invention thus provides a database containing genotype and phenotype information that can readily be used to obtain clinically and/or therapeutically and/or diagnostically useful information.

CLAIMS

1. A database comprising a plurality of records, each record containing genotype information, and optionally sample information for an individual, wherein:

the phenotype information for the individual comprises at least osteoporosis related phenotypes, osteoarthritis related phenotypes, immune cell subtypes (such as Tcell subsets), metabolic syndrome/syndrome X related phenotypes, and hypertension related phenotypes; and

the sample information for individual comprises information relating to the location of a sample of tissue or of fluid from the individual.

2. A database according to Claim 1, wherein the phenotype information further comprises thrombosis/fibrinolysis phenotypes, haemoglobinopathy related phenotypes and airways disease (asthma) phenotype.

3. A database according to Claim 1 or 2, wherein the phenotype information further comprises information relating to the phenotypes: atopy/eczema, lung function, IgE, psoriasis, acne, skin cancer and moliness of skin.

4. A database according to any preceding Claim comprising a plurality of records for human individuals.

5. A database according to any preceding Claim wherein the sample of tissue or of fluid is selected from the group consisting of urine, serum, skin, liver, heart, bone, hair, muscle, kidney, tooth, saliva, faeces and DNA.

6. A database according to any preceding Claim wherein the sample information comprises the geographical location of the sample, the storage conditions of the sample and the storage reference number for reference

label of the sample.

7. A database according to Claim 6 wherein the sample information additionally comprises contact information enabling the individual to be contacted and retested in person.

8. A database according to any preceding Claim, wherein each record further includes genotype information for the individual comprising one or more single nucleotide polymorphisms.

9. A database according to any of Claims 1 to 8, wherein the genotype information comprises information selected from one or more of:

(i) actual or inferred DNA base sequence at one or more regions within the genome;

(ii) a record of variation between a specified sequence on a chromosome of that individual compared to a reference sequence; and

(iii) length of a particular sequence or a particular sequence variant.

10. A method of adding information to a database according to any of Claims 1 - 9 comprising:

(1) identifying an individual not yet included in the database;

determining phenotype information for the individual;

optionally determining genotype information for the individual;

optionally determining sample information for the individual that includes information relating to the location of the sample of tissue or of fluid from the individual; and

creating a record in the database to hold the phenotype and optionally genotype and/or sample information for the individual;

or

(2) identifying an individual already included in a record in the database;

using sample information in the database to obtain a tissue or fluid sample for the individual;

testing the sample, thereby determining genotype or phenotype information for the individual; and

adding or confirming or amending or updating information in the record for the individual.

11. A method of identifying a correlation between phenotype information and genotype information comprising:

selecting a phenotype characteristic;

identifying a plurality of records from the database of the invention for individuals that comply with the selected phenotype characteristic;

determining if presence of the selected phenotype characteristic is correlated with presence of any genotype characteristic in the genotype information for records in the database.

12. A method of identifying a correlation between phenotype information and phenotype information comprising:

selecting a phenotype characteristic;

identifying a plurality of records in the database for individuals who comply with the phenotype characteristic;

determining if presence of the selected phenotype characteristic is correlated with another characteristic of phenotype information or records in the database.

13. A method of identifying a correlation between genotype information and genotype information comprising:

selecting a genotype characteristic;

identifying a plurality of records in the database for individuals who comply with the genotype characteristic;

determining if presence of the selected genotype characteristic is correlated with another characteristic of genotype information or records in the database.

14. A method of allocating priority to a candidate gene or locus, proposed as a drug target for treatment of a disease, the method comprising:-

calculating, from data on a database according to any of Claims 1 to 9, the specificity of the candidate gene or locus for the disease;

comparing (i) the association of the disease with clinical risk traits related to the disease, to (ii) the association of the disease with other clinical risk traits unrelated to the disease, but representing significant side effects; and

hence calculating a likely therapeutic index of drug candidates acting on that gene or locus.

15. A method of analysing the relation between a genotype and a phenotype, comprising

selecting a phenotype characteristic;

identifying a plurality of records in a database according to any of Claims 1 to 9 complying with that characteristic;

using environmental and age-related data in the database to eliminate the effects of age and environment on variations in phenotype; and

hence calculating from the database whether and if so to what extent the phenotype is correlated with a particular genotype.

16. A method of determining the capacity and specificity of a genetic marker to detect and quantify normal variations in healthy and affected populations for a selected risk trait, comprising:-

assaying samples in a database according to any of Claims 1 to 9 for the marker levels, in both healthy and affected subjects; and

quantifying the association of the clinical trait with the marker level and other selected phenotypes, in unaffected and affected subjects.

17. A method of devising dose regimes and/or dose forms and/or drug delivery systems for a given drug in a clinical trial, comprising:-

selecting a proposed clinical population for the trial;

using data on a database according to any of Claims 1 to 9 to stratify the clinical population by high associations of metabolism or absorption of the drug both with genotype and/or with associated biochemical and cell biology

phenotypes; and

hence allowing definition of the best dose regimes and dose forms/drug delivery systems;

so as to predict and/or allow for absorption and/or metabolism of the drug by patients in the clinical population.

18. A method of predicting response to a proposed drug therapy, comprising:-

using a database according to any of Claims 1 to 9 to select a clinical population by constructing haplotypic profiles, with strong associations with defined clinical traits and biochemical phenotypes;

using the database to eliminate the effects of age and environment in the clinical population;

hence providing criteria to predict response to the drug and variation in response to the drug, and optionally to define a sub-group of the clinical population or of the general population most susceptible to the drug being studied.

19. A database substantially as hereinbefore described and claimed in Claim 1.